Published online 13 November 2017 *Nucleic Acids Research, 2018, Vol. 46, Database issue* ***D497–D502***
doi: 10.1093/nar/gkx1130

An update on sORFs.org: a repository of small ORFs identified by ribosome profiling

Volodimir Olexiouk*, Wim Van Criekinge and Gerben Menschaert*

Lab of Bioinformatics and Computational Genomics (BioBix), Department of Mathematical Modelling, Statistics and Bioinformatics, Faculty of Bioscience Engineering, Ghent University, 9000 Ghent, Belgium

Received September 14, 2017; Revised October 25, 2017; Editorial Decision October 25, 2017; Accepted October 26, 2017

ABSTRACT

sORFs.org (http://www.sorfs.org) is a public repository of small open reading frames (sORFs) identified by ribosome profiling (RIBO-seq). This update elaborates on the major improvements implemented since its initial release. sORFs.org now additionally supports three more species (zebrafish, rat and *Caenorhabditis elegans*) and currently includes 78 RIBO-seq datasets, a vast increase compared to the three that were processed in the initial release. Therefore, a novel pipeline was constructed that also enables sORF detection in RIBO-seq datasets comprising solely elongating RIBO-seq data while previously, matching initiating RIBO-seq data was necessary to delineate the sORFs. Furthermore, a novel noise filtering algorithm was designed, able to distinguish sORFs with true ribosomal activity from simulated noise, consequently reducing the false positive identification rate. The inclusion of other species also led to the development of an inner BLAST pipeline, assessing sequence similarity between sORFs in the repository. Building on the proof of concept model in the initial release of sORFs.org, a full PRIDE-ReSpin pipeline was now released, reprocessing publicly available MS-based proteomics PRIDE datasets, reporting on true translation events. Next to reporting those identified peptides, sORFs.org allows visual inspection of the annotated spectra within the Lorikeet MS/MS viewer, thus enabling detailed manual inspection and interpretation.

INTRODUCTION

The probability of generating a start site ('ATG') by random sampling the nucleotide space is 1 out of 64. In addition, the probability of sampling a stop codon ('TAA', 'TAG', 'TGA') within the next 99 codons is ∼99%. Consequently, this implies that approximately 1.5% of the genome would consist of small open reading frames (sORFs, ≤300 nucleotides), assuming that the genome is generated by a random event, without considering splice events, reading frames, nucleotide biases, CG-content of the genome, or strandedness (1). Identifying translating sORFs in this vast pool of random sORFs is challenging, further complicated by the lack of sequence similarity between sORFs and known protein coding ORFs (2–4). Also, RNA-sequencing is unable to delineate ORFs and MS-based proteomic approaches have difficulties in detecting small protein products, illustrating the technological complications we are facing in the micropeptide detection process (5,6). As a result of this complex process, sORFs have historically been labelled as lacking coding potential. It is the advent of ribosome profiling (RIBO-seq) (7,8), that forced us to reconsider our opinion on the truly non-coding nature of these small ORFs (9–11).

Since the initial release of sORF.org, the Ensembl consortium (12) re-annotated 147 non-protein coding transcripts to protein coding (updated annotation from Ensembl version 81–90), where the protein product is <100 AA long. This set holds 54 long non-coding RNA (lncRNA) transcripts. Our initial release nourished this growing field on sORF-encoded polypeptides by establishing a first public portal bundling this focussed information, soon other initiatives followed, such as ARA-PEP (13) and SmProt (14). Here, an update on the sORFs.org repository is provided incorporating 78 new RIBO-seq datasets and including support for three new species, currently harbouring 34 human, 27 mouse, 5 rat, 3 zebrafish, 3 fruit fly and 6 *Caenorhabditis elegans* datasets. This vast increase in number of processed datasets (three at initial release) is mainly attributable to the development of a modified pipeline enabling the detection of sORFs in absence of data on initiating ribosomes, where an extra noise filtering step controls for false positive events.The addition of data on new species to sORFs.org drove the development of a 'between species' sORF BLAST (15) to detect sORFs with sequence similarity. Next, publicly available mass spectrometry (MS) datasets from PRIDE (16) are rescanned to acquire translational evidence for sORFs, as already available in our ini-

*To whom correspondence should be addressed. Tel: +32 9 264 99 22; Email: volodimir.olexiouk@ugent.be
Correspondence may also be addressed to Gerben Menschaert. Tel: +32 9 264 99 22; Email: gerben.menschaert@ugent.be

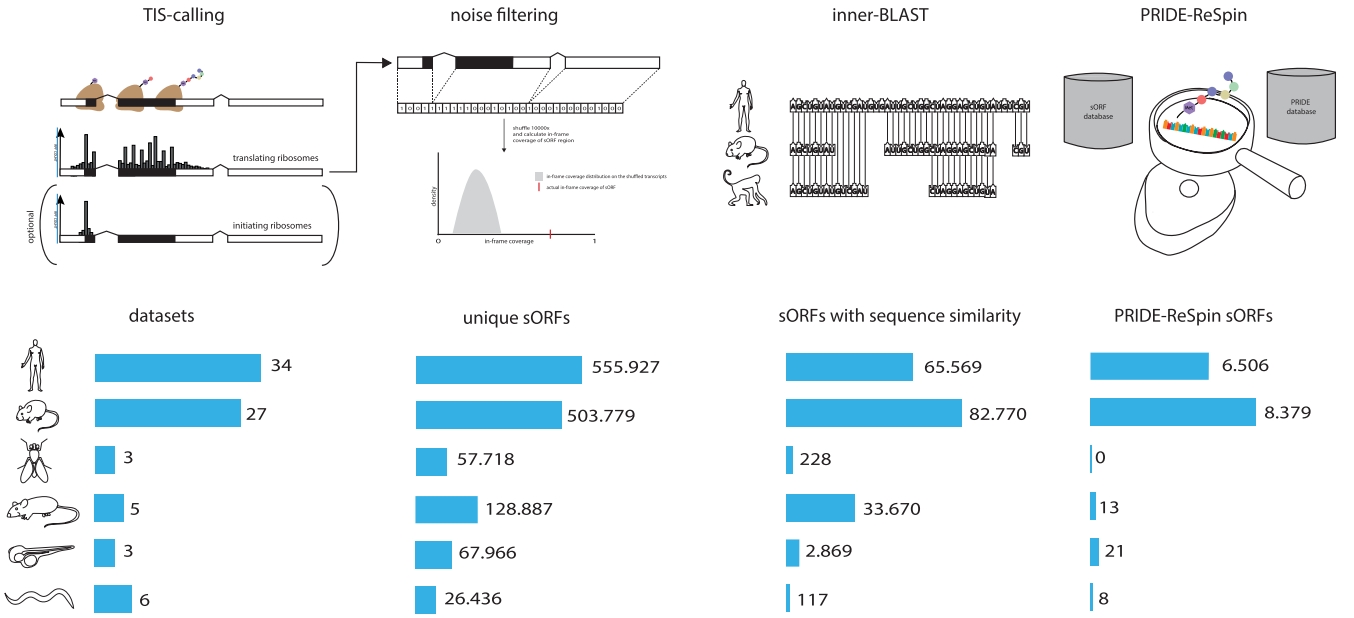

Figure 1. An overview of the most important improvements to sORFs.org since its initial release. The modified TIS-calling pipeline together with the noise filtering algorithm enabled the inclusion of datasets on additional species, wherefore no initiating RIBO-seq data (LTM or HAR treated) was available. Currently, a total of 78 RIBO-seq datasets are processed, identifying numerous novel sORFs with ribosome occupancy. Implementation of the inner-BLAST pipeline revealed sORFs with sequence similarity identified in multiple species and the PRIDE-ReSpin pipeline provides an extra layer of translation evidence based on MS data for a plethora of sORFs.

tial release of sORFs.org (17) as a proof of concept. Additionally, a visual platform was developed allowing the inspection of annotated identified MS/MS fragmentation spectra in the Lorikeet MS/MS viewer (https://github.com/jmchilton/lorikeet). This valuable feature provides a significant advantage over conventional MS-based identification reporting, which report identification either by a score, as in SmProt (14), or by a static figure (18). Figure 1 summarizes the most important improvements to sORFs.org since its initial release.

MATERIALS AND METHODS

Summary of the initial sORFs.org features

The initial release of sORFs.org provided 2 query interfaces. A default query interface enables quick, real-time lookup of specific sORFs whereas a second BioMart query interface (19) provides advanced query and export functionality. The query interfaces were optimized and improved based on community requests and input. Every sORF within the repository has its own detail page, bundling all available information. All metrics and information from our initial release (17) are still present, but we would like to stress that this page also contains two RIBO-seq coverage representations. A first one presents dataset-specific ribosome occupancy information within the UCSC genome browser interface (20), enabling inspection of the ribosome profile in or surrounding the sORF. A second intuitive in-house developed visualization allows more detailed inspection, allowing to select for certain reading frames or ribosome protected fragment (RPF) lengths. In our initial release, conservation was calculated using PhyloCSF (21), the inclusion of many new datasets constrained us to change to PhastCon (22) and PhyloP (23) due to computational limitation. However, in a future release we plan to optimize and implement PhyloCSF (21). Also, the BLASTp (15) search for sORFs against the non-redundant protein database from NCBI (24,25), which is periodically updated, is presented alongside.

TIS calling

The initial TIS-calling method required data on initiating ribosomes (e.g. by means of lactomidomycin (LTM) or harringtonine (HAR) treatment), with matching data on elongating ribosomes (e.g. by means of cycloheximide (CHX) treatment) (26). A limited amount of studies was published combining the two types of ribosome profiling experiments measuring both initiating and elongating ribosomes. This urged for the development of a modified TIS-calling algorithm based solely on translating ribosomes. In a first step, all start sites are identified genome-wide only taking into account the four most prominent start triplets 'ATG', 'CTG', 'TTG' and 'GTG', as opposed to the initial TIS-calling algorithm that considers all near cognate start triplets. Data on initiating ribosomes allows to pinpoint the correct TIS and the lack thereof increases the difficulty of non-ATG start site detection, resulting in an increase of truncations and extension caused by near-cognate start-sites occurring by chance. However, for well translated sORFs, data on initiating ribosomes should not be necessary for detection. Next, all start sites are scanned for an in-frame stop codon within 300nt, both with and without considering splice information extracted from the Ensembl annotation (12). For each possible sORF, the in-frame coverage and the RPF read count is calculated. A lenient threshold of at least

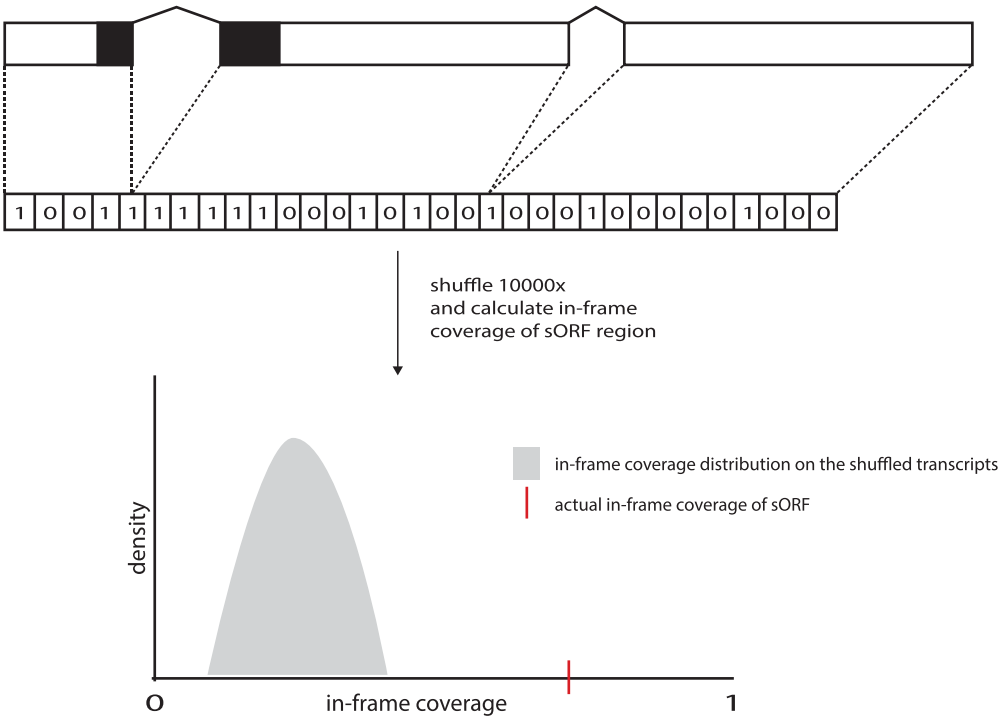

Figure 2. Visual representation of the noise filtering algorithm. The transcript of the sORF is reconstructed into a binary array, where '1' represent positions covered by ribosome P-site and '0' uncovered. This array is then shuffled 10 000 times, each iteration calculates the in-frame coverage in the sORF region, shaping a distribution of shuffled in-frame coverage as represented in gray. Next, the probability of sampling a value equal or greater than the actual in-frame coverage of the sORF is calculated (represented in red).

10% in-frame coverage and 10 RPFs is imposed to withhold sORFs. For those passing these criteria, the identified TIS are used in the assembly step as described in the initial release (17). The modified TIS-calling method enabled the addition of numerous datasets, resulting in the identification of novel sORFs as well as reoccurring sORFs (∼45% of sORFs are identified in multiple datasets, see supplementary file, Figure S1).

Noise filtering

As the novel TIS-calling algorithm does not build on two layers of evidence, comprising both data from elongating and initiating RIBO-seq experiments, it is clear that (non-AUG) start site prediction becomes more difficult and more false positive results are introduced. In order to counteract this, an accompanying novel noise filtering approach was developed comparing the RPF occupancy of sORFs with 'simulated' noise, trying to truly asses these translation events.

First, the transcript of the corresponding sORF is converted into a binary array, where '1' represents a position covered by ribosomes whereas '0' points to uncovered positions. After calculating the in-frame coverage for the sORF, this binary array is shuffled and the in-frame coverage is recalculated. This shuffling and recalculation of the in-frame coverage is repeated 10.000 times, creating a distribution of shuffled in-frame coverages, representing randomly allocated RPF coverage. Next, the probability is calculated to obtain an in-frame coverage of at least the actual in-frame coverage (Figure 2). The resulting *P*-values are subjected to the Benjamin–Hochberg (27) procedure for multiple testing to control the FDR at $\alpha = 0.05$. Notably, for intronic sORFs, the intron where the sORF resides is considered as an exon in the noise filtering step and for intergenic sORFs the transcript is considered to be the region 1000nt up- and down-stream of the sORF. Also, sORFs are inspected for overlap with any protein coding exon on any transcript, sORFs overlapping with protein coding exons are reported and sORFs overlapping and in-frame with the protein coding exons are discarded. The noise filtering algorithm has been validated on the crappé_2014 dataset (GSM1403307) using annotated canonical protein-coding transcript as a positive and 3′UTR regions as a negative control. These results are represented in supplementary Figures S4 and S5.

Inner BLAST

Addition of new species enabled us to investigate whether sORFs with sequence similarity over different species are present. Also, linking these related sORF sequences, provides experimentalists to perform functional characterization in a more convenient test model based on other organisms. The inner BLAST is performed by searching for sequence similarity in sORFs identified in distinct species using BLASTp (15) at an expected value of 0, 0000000001. Roughly 18% of sORFs express sequence similarity with at least one sORF (see supplementary file, Figure S2).

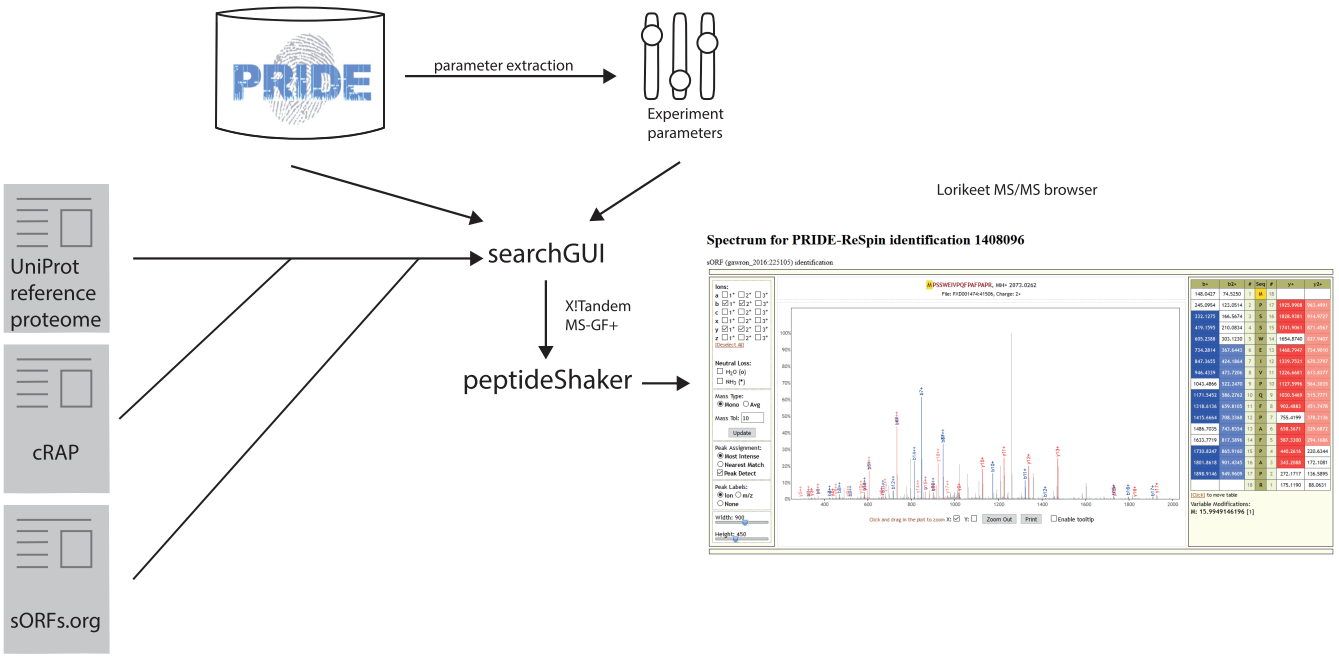

Figure 3. General overview of the PRIDE-ReSpin pipeline. First, MS-based proteomics experiments are downloaded from the PRIDE public repository. Next, a reverse engineering mechanism based on PRIDE-ASAP and Pladipus extracts the database (DB) search parameters for that study. These are inputted into the searchGUI search engine management software, launching a DB search against a concatenated database consisting of the UniProt reference proteome, the cRAP database and the sORFs.org database, using the X!Tandem and MS-GF+ as search engines. Consecutively, the output is imported into PeptideShaker to validate and export identified peptides at an FDR of 1%, with a minimum of 30% spectrum coverage and no PSMs having a higher confidence to non sORF peptides. These resulting peptides are then imported into sORFs.org for visualization in the Lorikeet MS/MS browser.

PRIDE-ReSpin

Acquiring proteomic evidence for micropeptides has proven to be strenuous (4,5,28,29). Many features such as their low abundance and putative hydrophobicity but also the lack of enzymatic cleavage sites and specific extraction protocols makes their identification hard with MS approaches. Yet, technological and computational advancements have recently resulted in the identification of several micropeptides using proteomics approaches (4,5,18,30,31). Including all possible translated sORF sequences on genome-wide scale impairs their identification and validation by inflating the search space, that is why these micropeptide sequences are generally excluded. sORFs.org provides a focussed database of putative micropeptides with translational evidence from RIBO-seq, suitable for inclusion into the search space within proteomics experiments. Most proteomics experiment are tailored for a specific purpose and are only examined once within the context of the study. Much more information thus remains undetected, which is gaining awareness in the community. The potential of reprocessing public proteomics datasets has been stressed (32–38), and is applied here for micropeptide detection.

The PRIDE-ReSpin runs continuously, periodically updating validated peptides to sORFs.org. At the time of writing, 302 human, 126 mouse, 18 rat, 10 zebrafish and 3 *C. elegans* datasets were processed identifying 463.678 PSMs that account for 10.583 uniquely identified peptides. For human, 291 3′-UTR, 675 5′-UTR, 1.954 exonic, 131 intronic, 129 intergenic, and 19 lincRNA unique sORF peptides were identified (see supplementary file, Figure S3). sORFs.org allows to visually inspect the identified spectra in the Lorikeet MS/MS browser, enabling manual assessment and validation of the identifications rather than bluntly reporting identified peptides (Figure 3). A detailed description of the PRIDE-ReSpin methodology can be found in the supplementary file.

COMPARISON WITH OTHER RESOURCES

Since the initial release of sORFs.org, several other public databases containing small open reading frame information emerged (39). The ARA-PEP repository (http://www.biw.kuleuven.be/CSB/ARA-PEPs/) (13) focusses on *Arabidopsis thaliana* and presents genomic, transcriptional and conservation information in order to annotate sORFs.

The smPROT repository (14) has more overlap with sORFs.org, harbouring a vast amount of identified sORFs across distinct species. smPROT uses the RiboTaper (40) tool to identify putatively translated sORFs from ribosome profiling data and thus significantly differs from our approach, which in not primarily based on the triplet periodicity. sORFs.org includes all sORFs with evidence of ribosome occupancy and computes various sORF translation detection metrics (e.g. FLOSS, ORFscore) alongside genomic and proteomic features, thus providing researchers the capability to tailor sORFs.org information to their own research projects, using our query interfaces. SmProt provides limited translation detection metrics and genomic features (conservation, variation), however, detects sORFs from literature mining, a feature currently missing in sORFs.org. Furthermore, sORFs.org aims to be as trans-

parent as possible in data acquisition and processing, providing information and statistics both on the datasets used, as well as providing visual tools to inspect data, for instance by representing RPF data in the UCSC genome browser (20) or in our in-house developed browser. In contrast, smPROT reports only limited genomic and RPF based features and provides no means to inspect the credibility of the reported information. This in our opinion is a very important feature, especially in this field where false positive detection is possible. Also, smPROT reports 117.099 sORFs with MS-evidence including 83.159 exonic sORFs, 24.539 lincRNA sORFs, 5.272 antisense sORFs and 1.854 'sense no exonic' sORFs. This huge amount of identified micropeptides based on MS information has not been corroborated by us or other studies. As the smPROT does not have the ability to validate/inspect the MS data—only a raw score of the identified peptide is reported—these findings could not be verified. sORFs.org allows the inspection of matched fragmentation through the Lorikeet viewer and also dynamically scans more deposited dataset based on the PRIDE-ReSpin approach, which is in shear contrast to the smPROT database.

CONCLUSION AND FUTURE PERSPECTIVES

sORFs now additionally supports three species (rat, zebrafish and *C. elegans*) and includes 78 extra datasets. This has been achieved by implementing a novel TIS-calling algorithm, enabling the identification of sORFs from RIBO-seq experiments comprising solely elongating ribosome data (through CHX treatment). Moreover, a novel noise filtering algorithm was devised to distinguish sORFs translation events with true ribosome occupancy from simulated noise. The addition of new species led to the development of the inner-BLAST pipeline, identifying homologues sORFs in our repository. Lastly, the PRIDE-ReSpin MS data reprocessing pipeline was released and incorporated into sORFs.org, periodically scanning publicly available datasets to acquire relevant translational evidence for sORFs. The Lorikeet MS/MS viewer ensures visual inspection of the annotated fragmentation spectra.

sORFs.org will continue to periodically include new datasets supporting extra species. Also, the PRIDE-ReSpin will be fine-tuned and optimized, increasing the amount of processable data. To build in a second layer of translational evidence based on MS, integration of sORFs.org with PeptideAtlas (41) and NextProt (42) is investigated. At present, the incorporation of small linear motives (sLIM) into sORFs.org is examined, by exploring the potential integration with the ELM database (43). Also, ways to incorporate protein family domains and motives such as pFAM (44) are investigated (including e.g. of transmembrane motives (45) and signal peptides (46)). In general, integration with different sources such as HaltORF (47) and RPFdb (48) will strengthen sORFs.org by accumulating relevant evidence for translation. A text-mining approach could help the annotation of sORFs by reporting recent scientific manuscripts. In all, sORFs.org continuously will follow the sORF research community enabling the implementation of novel features when requested.

AVAILABILITY

sORF.org is publicly available at http://www.sorfs.org. The underlying pipelines used for sORFs.org can be made available upon request, however, were not optimized for public usage.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We want acknowledge Kenneth Verheggen and Lennart Martens for their contribution in creating and improving the parameter extraction pipeline for PRIDE-ReSpin.

FUNDING

Postdoctoral Fellows of the Research Foundation – Flanders (FWO-Vlaanderen) [12A7813N to G.M.]; Research Foundation—Flanders (FWO-Vlaanderen) [G0D3114N to V.O.]. Funding for open access charge: Ghent University

Conflict of interest statement. None declared.

REFERENCES

1. Pauli,A., Valen,E. and Schier,A.F. (2015) Identifying (non-)coding RNAs and small peptides: challenges and opportunities. *Bioessays*, **37**, 103–112.
2. Makarewich,C.A. and Olson,E.N. (2017) Mining for micropeptides. *Trends Cell Biol.*, doi:10.1016/j.tcb.2017.04.006.
3. Couso,J.-P. and Patraquim,P. (2017) Classification and function of small open reading frames. *Nat. Rev. Mol. Cell Biol.*, doi:10.1038/nrm.2017.58.
4. Saghatelian,A. and Couso,J.P. (2015) Discovery and characterization of smORF-encoded bioactive polypeptides. *Nat. Chem. Biol.*, **11**, 909–916.
5. Olexiouk,V. and Menschaert,G. (2016) Identification of small novel coding sequences, a proteogenomics endeavor. In: *Advances in Experimental Medicine and Biology*.Vol. **926**, pp. 49–64.
6. Yagoub,D., Tay,A.P., Chen,Z., Hamey,J.J., Cai,C., Chia,S.Z., Hart-Smith,G. and Wilkins,M.R. (2015) Proteogenomic discovery of a small, novel protein in yeast reveals a strategy for the detection of unannotated short open reading frames. *J. Proteome Res.*, doi:10.1021/acs.jproteome.5b00734.
7. Ingolia,N.T., Ghaemmaghami,S., Newman,J.R.S. and Weissman,J.S. (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, **324**, 218–223.
8. Ingolia,N.T., Brar,G.A., Rouskin,S., McGeachy,A.M. and Weissman,J.S. (2012) The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat. Protoc.*, **7**, 1534–1550.
9. Bazzini,A.A., Johnstone,T.G., Christiano,R., MacKowiak,S.D., Obermayer,B., Fleming,E.S., Vejnar,C.E., Lee,M.T., Rajewsky,N., Walther,T.C. *et al.* (2014) Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J.*, **33**, 981–993.
10. Ingolia,N.T., Lareau,L.F. and Weissman,J.S. (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, **147**, 789–802.
11. Ingolia,N.T., Brar,G.A., Stern-Ginossar,N., Harris,M.S., Talhouarne,G.J.S., Jackson,S.E., Wills,M.R. and Weissman,J.S. (2014) Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep.*, doi:10.1016/j.celrep.2014.07.045.
12. Yates,A., Akanni,W., Amode,M.R., Barrell,D., Billis,K., Carvalho-Silva,D., Cummins,C., Clapham,P., Fitzgerald,S., Gil,L. *et al.* (2016) Ensembl 2016. *Nucleic Acids Res.*, **44**, D710–D716.

13. Hazarika,R.R., De Coninck,B., Yamamoto,L.R., Martin,L.R., Cammue,B.P.A. and van Noort,V. (2017) ARA-PEPs: a repository of putative sORF-encoded peptides in Arabidopsis thaliana. *BMC Bioinformatics*, **18**, 37.
14. Hao,Y., Zhang,L., Niu,Y., Cai,T., Luo,J., He,S., Zhang,B., Zhang,D., Qin,Y., Yang,F. *et al.* (2017) SmProt: a database of small proteins encoded by annotated coding and non-coding RNA loci. *Brief. Bioinform.*, doi:10.1093/bib/bbx005.
15. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
16. Martens,L., Hermjakob,H., Jones,P., Adamsk,M., Taylor,C., States,D., Gevaert,K., Vandekerckhove,J. and Apweiler,R. (2005) PRIDE: the proteomics identifications database. *Proteomics*, **5**, 3537–3545.
17. Olexiouk,V., Crappé,J., Verbruggen,S., Verhegen,K., Martens,L. and Menschaert,G. (2015) sORFs.org: a repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res.*, doi:10.1093/nar/gkv1175.
18. Mackowiak,S.D., Zauber,H., Bielow,C., Thiel,D., Kutz,K., Calviello,L., Mastrobuoni,G., Rajewsky,N., Kempa,S., Selbach,M. *et al.* (2015) Extensive identification and analysis of conserved small ORFs in animals. *Genome Biol.*, **16**, 179.
19. Smedley,D., Haider,S., Durinck,S., Pandini,L., Provero,P., Allen,J., Arnaiz,O., Awedh,M.H., Baldock,R., Barbiera,G. *et al.* (2015) The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res.*, **43**, 589–598.
20. Rosenbloom,K.R., Armstrong,J., Barber,G.P., Casper,J., Clawson,H., Diekhans,M., Dreszer,T.R., Fujita,P.A., Guruvadoo,L., Haeussler,M. *et al.* (2015) The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.*, **43**, D670–D681.
21. Lin,M.F., Jungreis,I. and Kellis,M. (2011) PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, **27**, 275–282.
22. Siepel,A. and Haussler,D. (2005) Phylogenetic Hidden Markov Models. *Engineering*, doi:10.1089/1066527041410472.
23. Pollard,K.S., Hubisz,M.J., Rosenbloom,K.R. and Siepel,A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, **20**, 110–121.
24. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*., **44**, D733–D745.
25. Ostell,J. and McEntyre,J. (2007) *The NCBI Handbook*. NCBI Bookshelf.
26. Ingolia,N.T., Ghaemmaghami,S., Newman,J.R.S. and Weissman,J.S. (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, **324**, 218–223.
27. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.
28. Slavoff,S.A., Mitchell,A.J., Schwaid,A.G., Cabili,M.N., Ma,J., Levin,J.Z., Karger,A.D., Budnik,B.A., Rinn,J.L. and Saghatelian,A. (2013) Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat. Chem. Biol.*, **9**, 59–64.
29. Yang,X., Tschaplinski,T.J., Hurst,G.B., Jawdy,S., Abraham,P.E., Lankford,P.K., Adams,R.M., Shah,M.B., Hettich,R.L., Lindquist,E. *et al.* (2011) Discovery and annotation of small proteins using genomics, proteomics, and computational approaches. *Genome Res.*, **21**, 634–641.
30. Ma,J., Ward,C.C., Jungreis,I., Slavoff,S.A., Schwaid,A.G., Neveu,J., Budnik,B.A., Kellis,M. and Saghatelian,A. (2014) Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue. *J. Proteome Res.*, **13**, 1757–1765.
31. Chu,Q., Ma,J. and Saghatelian,A. (2015) Identification and characterization of sORF-encoded polypeptides. *Crit. Rev. Biochem. Mol. Biol.*, **50**, 134–141.
32. Vaudel,M., Verheggen,K., Csordas,A., Ræder,H. and Frode,S. (2015) Exploring the potential of public proteomics data. *Proteomics*, doi:10.1002/pmic.201500295.
33. Rung,J. and Brazma,A. (2013) Reuse of public genome-wide gene expression data. *Nat. Rev. Genet.*, **14**, 89–99.
34. Perez-Riverol,Y., Xu,Q.W., Wang,R., Uszkoreit,J., Griss,J., Sanchez,A., Reisinger,F., Csordas,A., Ternent,T., Del-Toro,N. *et al.* (2016) PRIDE Inspector Toolsuite: moving towards a universal visualization tool for proteomics data standard formats and quality assessment of ProteomeXchange datasets. *Mol. Cell. Proteomics*, **15**, 305–317.
35. Vaudel,M., Burkhart,J.M., Zahedi,R.P., Oveland,E., Berven,F.S., Sickmann,A., Martens,L. and Barsnes,H. (2015) PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat. Biotechnol.*, **33**, 22–24.
36. Sevinsky,J.R., Cargile,B.J., Bunger,M.K., Meng,F., Yates,N.A., Hendrickson,R.C. and Stephenson,J.L. (2008) Whole genome searching with shotgun proteomic data: applications for genome annotation. *J. Proteome Res.*, **7**, 80–88.
37. Perez-Riverol,Y., Alpi,E., Wang,R., Hermjakob,H. and Vizcaíno,J.A. (2015) Making proteomics data accessible and reusable: Current state of proteomics databases and repositories. *Proteomics*, **15**, 930–950.
38. Martens,L. and Vizcaíno,J.A. (2017) A golden age for working with public proteomics data. *Trends Biochem. Sci.*, **42**, 333–341.
39. Plaza,S., Menschaert,G. and Payre,F. (2017) In search of lost small peptides. *Annu. Rev. Cell Dev. Biol.*, **33**, 391–416.
40. Calviello,L., Mukherjee,N., Wyler,E., Zauber,H., Hirsekorn,A., Selbach,M., Landthaler,M., Obermayer,B. and Ohler,U. (2015) Detecting actively translated open reading frames in ribosome profiling data. *Nat. Methods*, doi:10.1038/nmeth.3688.
41. Deutsch,E.W., Sun,Z., Campbell,D., Kusebauch,U., Chu,C.S., Mendoza,L., Shteynberg,D., Omenn,G.S. and Moritz,R.L. (2015) State of the human proteome in 2014/2015 As viewed through peptideatlas: Enhancing accuracy and coverage through the atlas prophet. *J. Proteome Res.*, **14**, 3461–3473.
42. Gaudet,P., Michel,P.A., Zahn-Zabal,M., Britan,A., Cusin,I., Domagalski,M., Duek,P.D., Gateau,A., Gleizes,A., Hinard,V. *et al.* (2017) The neXtProt knowledgebase on human proteins: 2017 update. *Nucleic Acids Res.*, **45**, D177–D182.
43. Dinkel,H., Roey,K., Michael,S., Kumar,M., Uyar,B., Altenberg,B., Milchevskaya,V., Schneider,M., Kühn,H., Behrendt,A. *et al.* (2016) ELM 2016—data update and new functionality of the eukaryotic linear motif resource. *Nucleic Acids Res.*, **44**, D294–D300.
44. Finn,R.D., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Mistry,J., Mitchell,A.L., Potter,S.C., Punta,M., Qureshi,M., Sangrador-Vegas,A. *et al.* (2016) The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.
45. Mulay,S.R., Desai,J., Kumar,S.V., Eberhard,J.N., Thomasova,D., Romoli,S., Grigorescu,M., Kulkarni,O.P., Popper,B., Vielhauer,V. *et al.* (2016) Secreted protein prediction system combining CJ-SPHMM, TMHMM, and PSORT. *Nat. Commun.*, **7**, 10274.
46. Petersen,T.N., Brunak,S., von Heijne,G. and Nielsen,H. (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods*, **8**, 785–786.
47. Vanderperre,B., Lucier,J.F. and Roucou,X. (2012) HAltORF: Adatabase of predicted out-of-frame alternative open reading frames in human. *Database*, **2012**, bas025.
48. Xie,S.-Q., Nie,P., Wang,Y., Wang,H., Li,H., Yang,Z., Liu,Y., Ren,J. and Xie,Z. (2015) RPFdb: a database for genome wide information of translated mRNA generated from ribosome profiling. *Nucleic Acids Res.*, doi:10.1093/nar/gkv972.